



Machine Readable Archives

BULLETIN

ISSN 0821-3658

A Union List of Canadian Machine Readable Data Files: Past, Present and Future

On reviewing the history of the development of data archives and libraries in Canada, two major concerns have been raised continually since the late sixties: firstly, that the documentation or codebook for the data is complete and that standard coding formats are used; and secondly, that a comprehensive inventory of machine readable data files is made available for secondary research. Attempts have been made to address the creation of an inventory; however, maintenance of the inventory has not been possible owing to lack of permanent funding. The importance of such an inventory has been of continuing interest to both the data archival community and the research community who use social science data. Many researchers must use American data, as knowledge of Canadian data does not readily exist. Lack of information on sources of machine readable data in Canada can result in duplication of effort and expenditure.

Use of computers for statistical purposes began in the sixties. Both governments and university social science programs began to use computers for the creation and manipulation of socio-economic data. Statistics Canada was, and still is, the major producer of such data. Other federal departments, however, have also become major producers either in-house or through the use of contracts. The Royal Commission on Bilingualism and Biculturalism produced several national surveys through contracts with social science researchers. Universities began to create data archives to acquire, store and disseminate data sets to academics. The Institute for Behavioural Research maintained an advanced survey research centre, which in 1971 was organized into three specialized units: the Canadian Attitude Archive, the Cross-national Archive and the Census Data Archive. The University of Saskatchewan maintained the Sample Survey Centre and Data Bank and the University of Montreal operated the Centre de Sondage. Data libraries and archives were small but active centres for the dissemination of machine readable data.

The concern over the lack of information about existing Canadian sources of data was formalized in the early seventies in a report prepared for the Joint Steering Committee of the Association of Universities and Colleges of Canada (AUCC) and the Social Science Research Council of Canada (SSRCC). Mr. H.C. Campbell, Chief Librarian of the Toronto Public Libraries at the time, under-

took a study, "Data Clearing House for the Social Sciences," in which he reviewed the state of machine readable data in the early seventies. Also included in the study was a cursory inventory of holdings of various data centres and a survey of one hundred and fifty-seven faculty users and producers of research data from Canadian universities. Although not all of the faculty members were using or creating machine readable data at that time, most predicted fairly substantial use of such data over the next five years. In his report, Mr. Campbell concluded that the creation and use of machine readable data was evident in all provinces and in all sectors. He also acknowledged the large number of research grants awarded by the Canada Council to the social sciences, many of which produced quantitative data. His recommendation, which was accepted by the Joint AUCC/SSRCC Steering Committee, was the creation of a social science data clearing house. The major objectives of the clearing house were to be: (1) the preparation of an index of social data holdings in machine readable form held by Canadian universities, non-profit research agencies and other bodies involved in social science research; (2) the collection of descriptions of machine readable data from federal and provincial governments and the operation as a liaison between individual researchers and government departments; (3) the provision of information to individual researchers on the location of data; and (4) the provision of technical information for effective use of the data.

Based on this recommendation, the Data Clearing House for the Social Sciences (DCHSS) became a reality. One of its first major tasks was to prepare a standard description form to be used in the collection of information on machine readable data files. DCHSS encouraged the input of ideas from the data archival community in Canada on the elements to be included in the collection of descriptive information. Over the next few years, DCHSS was able to collect valuable information on the location and contents of machine readable data in existence in the university sector, the private sector and governments at all levels. A major database was created and a useful published inventory was distributed. The DCHSS also published a bulletin that contained information on archival activities, national and international; major research projects; new sources of data; and other activities of interest

to the data user community. Unfortunately, the DCHSS was not able to obtain funding to continue its work. This was a major loss to the social science community as the DCHSS had been most successful in obtaining descriptive information from government sources. Another major loss was that of the inventory in its computerized form. The tapes have never been found and therefore no other organization was able to take over the updating of the inventory.

Since the demise of the DCHSS in 1979, the Machine Readable Archives has undertaken two projects that have revealed that the need for an inventory is still very important to the social science community. Results of the User Survey undertaken in 1982 indicated that an inventory or union list of machine readable data files, whether on-line or in published form, was by far the most needed tool. At the Vancouver Consultation in June 1983, and throughout the interviews that lead to the report on the "Canadian Electronic Cultural Heritage," the need for an inventory was expressed many times. Some researchers felt that such an inventory should be undertaken by an organization that had permanent funding and that perhaps it could be done on a smaller scale. Many specialized inventories have been completed over the past few years; however, each has been limited in its scope compared to what the DCHSS attempted to do.

Since funding for social science research has diminished in the eighties, it is important to reduce research costs through the knowledge and use of existing records. A tremendous amount of data is being produced that could be used, if identified.

Standardization of descriptive elements for machine readable data have been defined and information management packages are now more flexible in accepting data in other formats. These facts now make possible the creation of a less costly and more useful inventory. The MRA has been seriously considering a pilot project to determine the cost and feasibility of creating such an inventory. The pilot project would include approximately ten data archives that would submit descriptive information about their holdings to one centre. This centre would then convert the data into a standard format; or if the descriptive information was not in machine readable form the centre would do the data entry. A printed inventory would be available. The first task in the pilot project would be to reach agreement on the descriptive elements to be used. It would then be necessary to determine the feasibility and cost of the inventory. The Machine Readable Archives is planning to fund a



"A Union List . . ."

cont'd from page 1

meeting of a limited number of data archivists and librarians in the late fall. The purpose of the meeting will be to determine the descriptive elements that are to be used and to decide on the next steps involved. It is hoped that a Union List of Canadian Machine Readable Data Files will be a reality within the next two years.

IASSIST '84

Over ninety data archivists, librarians, researchers and other users and creators of machine readable data attended the IASSIST '84 conference in Ottawa, May 15 to May 18. The week began with a day of pre-conference workshops designed to provide practical information and experience on specific aspects of machine readable data. Workshop topics included: Complex Data Files, Microcomputers, Data Library Management, and Cataloguing Microcomputer Data Files.

The conference opened with an international panel of representatives of statistical agencies from Canada, Sweden, England and the United States. The topic of the panel was "Issues Confronting Statistical Agencies: An International Perspective." Members of the panel included: Christopher Denham, Office of Population Census and Surveys, England; Edmund Rapaport, Statistics Sweden; Lorne Rowebottom, Statistics Canada; and Paul Zeisset, Bureau of the Census, United States. Each representative outlined the activities of his respective agency addressing specific problems that are and will be confronted in the next few years. Similar issues were addressed and the audience was able to see how different countries were confronting similar issues. The session was very informative and stimulated many questions from both the panel and the audience.

The sub-theme of Privacy and Confidentiality began in the afternoon with a plenary session focusing on the researcher's view of these issues. Professor John Bossons of the Institute for Political Analysis, University of Toronto, addressed the problems experienced in obtaining and using microdata. Thomas Brown, Vice-President of the Association of Public Data Users and David Worton of Statistics Canada, responded to the issues raised by Professor Bossons.

The plenary session lead into three concurrent sessions, each dealing with a specific aspect of privacy and confidentiality: anonymization techniques, which addressed techniques used by statistical agencies and private companies in their collection and use of data; legislative aspects associated with machine

readable data in an international perspective; and the implications on archives of privacy and confidentiality issues, in particular the effect on distribution and standardization.

The second sub-theme, the Advance of Technology, addressed the advances in statistical software packages and their adaptation to microcomputers. Nancy Morisson, SPSS Inc. and Gary Anderson, SIR/DBMS Inc. were the major speakers and provided insights into the changes being made in these packages. It was suggested that a useful addition to this type of session would be a user of these packages who could discuss problems encountered in using the software.

Three concurrent sessions followed, each one dealing with a specific aspect of the topic. "To Minis and Micros" dealt with the adaptations or approaches required by archivists and researchers in making the transition from mainframes to minis and micros. The "Complex Data Bases" session addressed the problems experienced by researchers in analysing data in hierarchical and relational data bases. The "On-Line Bibliographic Systems" session discussed the different approaches that can be taken in developing on-line bibliographic systems.

The third sub-theme, Changing Roles and Responsibilities, began with a session on "The Growth of the Information Elite." Nancy Brodie, National Library of Canada; Joseph Paradi, Dataline Systems Limited; and Erika Von Brunken, Karolinska Institute Library and Information Centre, Sweden, debated the question of whether a new elite was being created due to the cost of equipment and knowledge of the use of computers. Three concurrent sessions followed: "Data Archives and Libraries: New Challenges," which assessed the impact of government policies on the establishment and maintenance of data archives; "Data Collection and Use: New Directions," which addressed new trends in data processing and their impact on governments, universities and private companies; and finally, the "Collection and Use of International Data," which outlined the problems that the collection of data from cross-national sources poses to both archivists and researchers.

Lunch hour meetings were held for two groups: DataLink and the IASSIST Working Group on the Preparation of an International Standard Bibliographic Description for MRDF. The latter met to discuss a draft working paper and approve recommendations on changes to specific cataloguing rules. This report will be submitted to the ISBD(NBM) Review Group.

Discussions carried over into the social activities of the conference. Both Statistics Canada and the Public Archives sponsored

receptions for participants. A Tenth Anniversary buffet/reception was held on the last night, followed by a digital disco. A tour of the city in a double-decker bus, with a stop at Statistics Canada to see a CANSIM demonstration, was organized. The week provided an excellent opportunity for participants to meet with colleagues from North America and Europe and discuss issues and problems that are confronting the profession.

The Program and Local Arrangements Committees would like to thank all participants for their enthusiastic participation in the conference. Next year's conference will be in Amsterdam in May. The theme of the conference is "Public Access to Public Data."

Computer Conferencing among Canadian Data Archives and Libraries

In August 1983, Wendy Watkins, Data Archivist at Carleton University; Charles Humphrey, Data Library Analyst at the University of Alberta; and Laine Ruus, Data Librarian at the University of British Columbia, submitted to the Social Science and Humanities Research Council a proposal for the establishment of a communication link among Canadian data archives and libraries using computer conferencing technology. The objective was to improve communication among the repositories, and the proposal requested funds to support use of the system. Funding for a one-year trial was provided. Mr. Humphrey contacted as many archives and libraries as possible, inviting their participation. Approximately twenty institutions responded. A session was held at the IASSIST '84 conference to introduce the conferencing system and discuss its application.

Not only is the conferencing system being used to improve communication, but it also has as a major objective the development of a set of recommended standards for the preparation of machine readable data files. Through the conferencing system all users will be able to contribute their ideas and suggestions on the development of these standards.

The data archives and library community is a small one, both in numbers and in funding. Efforts have been made to organize meetings to discuss issues but not everyone has been able to attend. It is hoped that the conferencing system will provide a low-cost way to improve communications and to discuss and resolve problems associated with the archiving of machine readable data files.

The articles in this issue were prepared by Sue Gavrel, MRA.

de l'Inventaire. Les Archives ordinolingues prévoient financer la tenue d'une réunion d'un nombre restreint d'archivistes et de bibliothécares à la fin de l'automne. La réunion aurait pour but de choisir les éléments descriptifs et de fixer les prochaines étapes du projet. Nous espérons qu'un catalogue collectif de fichiers de données ordinolingues canadiens puisse voir le jour d'ici environ deux ans.

IASSIST 84

Plus de quatre-vingt-dix archivistes, bibliothécaires, chercheurs et autres producteurs et utilisateurs de données ordinolingues ont assisté à la conférence de l'IASSIST 84, qui a eu lieu du 15 au 18 mai, à Ottawa. La semaine a commencé par une journée préparatoire d'information et d'ateliers pratiques sur des aspects particuliers des données ordinolingues. Les thèmes des ateliers étaient les suivants : fichiers de données complexes, micro-ordinateurs, gestion des bibliothèques, et catalogue des fichiers de données générées par micro-ordinateur.

La conférence a débuté par une table ronde de représentants de bureaux de statistiques du Canada, de la Suède, de l'Angleterre et des États-Unis. Le thème de la discussion était « Questions d'intérêt pour les bureaux de statistique dans un contexte international ». La table ronde était constituée de M. Christopher Denham, Office of Population Census and Surveys (Angleterre); M. Edmund Rapaport, Statistisk Sued; M. Lorne Rowebottom, Statistics Canada; et M. Paul Zeisser, Bureau of the Census (États-Unis). Chaque représentant a donné un aperçu des activités de son bureau, en soulignant les difficultés à surmonter précédemment et dans les années à venir. Les participants ont pu constater de quelle façon les différents pays faisaient face aux mêmes difficultés. La séance a été très enrichissante et il y a eu une participation active de la part des conférenciers et de l'auditoire.

Le sous-thème sur les renseignements personnels et confidentiels a été abordé l'après-midi, la séance plénière étant axée sur l'opinion des chercheurs à cet égard. Le professeur John Bossons de l'Institute for Political Analysis de l'université de Toronto a souligné le problème que posent l'obtention et l'utilisation des micro-données. M. Thomas Brown, vice-président de l'Association of Public Data Users, et M. David Worton, de Statistique Canada, ont répondu à la question du professeur Bossons.

Après la séance plénière, les participants se sont divisés en trois groupes pour étudier plus à fond un aspect particulier de la protection des renseignements personnels : les techniques d'anonymat utilisées par les bureaux de statistique et les sociétés privées pour la collection et l'utilisation des données; les aspects législatifs des données ordinolingues dans un contexte international; et les répercussions sur les archives des questions de protection des

renseignements personnels et confidentiels, particulièrement l'incidence sur la diffusion et la normalisation.

Le deuxième sous-thème, l'évolution technologique, traitait des réalisations en ce qui concerne les programmes statistiques des logiciels et de leur adaptation aux micro-ordinateurs. Les principaux orateurs, Nancy Morisson, de SPSS Inc., et Gary Anderson, de SIR/DBMS Inc., ont expliqué les changements aux programmes. Il a été proposé d'inviter à l'avenir un utilisateur des programmes pour qu'il puisse discuter des difficultés que soulève le logiciel.

Trois séances concomitantes ont suivi, chacune traitant d'un aspect particulier de la question. « Des mini et des micro » portait sur l'adaptation des archivistes et des chercheurs aux mini et micro-ordinateurs. La séance sur « les bases de données complexes » abordait les difficultés qu'ont les chercheurs à analyser les données dans les bases de données hiérarchiques et relationnelles. La séance intitulée « systèmes bibliographiques en direct » examinait les différentes démarches à suivre pour établir des systèmes bibliographiques en direct.

Le troisième sous-thème, changement de rôle et d'attributions, a été abordé dans une séance sur « l'émergence d'une élite de l'information ». Nancy Brodie, Bibliothèque nationale du Canada, Joseph Parad, Dataline Systems Limited, et Erika Von Brunken, Bibliothèque et Centre d'information de l'Institut Karolinska, en Suède, ont discuté de la question à savoir si une nouvelle élite était en train de se former à cause du coût du matériel et des connaissances qu'exige l'utilisation des ordinateurs. Trois séances concomitantes ont suivi : « les nouveaux défis pour les archives et les bibliothèques » évaluaient l'incidence des politiques gouvernementales sur la création et l'entretien des archives; « la collection et l'utilisation des données internationales » cernait les problèmes que pose la collecte des données de sources internationales tant aux archivistes qu'aux chercheurs.

À l'heure du déjeuner, deux groupes se réunissaient : Datalink et le groupe de travail sur la description bibliographique internationale nolingues (MIRDF) de l'IASSIST. Ce dernier s'est réuni pour étudier un document de travail provisoire et adopter les recommandations concernant les changements à certaines règles de catalogage. Le rapport sera présenté au groupe de révision de l'ISBD (NBM).

Les discussions se sont poursuivies dans le cadre des activités sociales de la conférence. L'honneur des participants. Ceux-ci ont été ont été les hôtes des réceptions données en Statistique Canada et les Archives publiques.

Invités, le dernier soir, à un buffet dansant. Une visite guidée de la ville avec arrêt à Statistique Canada pour assister à une séance de démonstration CANSIM a été organisée. L'occasion de rencontrer des collègues nord-américains et européens et de discuter de sujets communs.

Les comités du programme et de l'organisation matérielle remercient tous les participants de l'enthousiasme qu'ils ont manifesté au cours de la conférence. La prochaine conférence aura lieu au mois de mai prochain à Amsterdam, sous le thème « l'accès du grand public aux données ».

Télécommunications par ordinateur dans les archives et les bibliothèques

En août 1983, M^{me} Wendy Watkins, archiviste des données à l'université Carleton, M. Charles Humphrey, analyste des données de bibliothèque à l'université d'Alberta, et Laine Ruus, bibliothécaire des données à l'université de Colombie-Britannique, ont présenté au Conseil de recherches en sciences humaines une proposition concernant l'établissement d'un système de télécommunications entre les archives et les bibliothèques, assisté par ordinateur. L'objectif était d'améliorer les communications entre les divers dépôts, et la proposition était appuyée d'une demande de crédits pour permettre l'exploitation du système. Des subsides ont été accordés pour une mise à l'essai d'un an. M. Humphrey a invité le plus grand nombre d'archives et de bibliothèques possible à participer à l'essai. Une vingtaine d'établissements ont répondu à l'invitation. Une séance a été organisée dans le cadre de la conférence de l'IASSIST 84 pour présenter le nouveau système et discuter de ses applications. Non seulement ce système améliore-t-il les communications, mais il permet l'établissement d'une série de normes relatives à la conservation des fichiers de données ordinolingues. En outre, les usagers pourront donner leur avis et faire des suggestions sur la création des normes.

Les archivistes des données et les bibliothécaires des données sont peu nombreux, et leurs ressources financières sont faibles. Des efforts ont été faits en vue d'organiser des réunions, mais la participation était relativement restreinte. Nous espérons que le système de téléinformatique permettra d'améliorer les communications de façon économique, ainsi que d'étudier et de résoudre les problèmes que pose l'archivage des fichiers de données ordinolingues.

Les articles ont été rédigés par Sue Gavel, DAO.



Archives ordinolinguas



Historique d'un catalogue collectif de fichiers de données ordinales

de données ordinoliques

de recherche, les nouvelles sources de données, et d'autres sujets d'intérêt pour les usagers. Malheureusement, le Centre ne put obtenir les crédits nécessaires pour poursuivre ses travaux. Ce fut une grande perte pour le milieu des sciences sociales, d'autant plus qu'il avait réussi à obtenir des renseignements décrits automatiquement et enregistré. Les bandes n'ayant jamais été retrouvées, aucun autre organisme ne put procéder à sa mise à jour. Depuis la disparition du Centre d'échange de données en sciences sociales, en 1979, les Archives ordinologiques ont mené deux enquêtes qui ont réaffirmé la nécessité de dresser un inventaire à l'intention du milieu des sciences sociales. Les résultats de l'Enquête sur les usagers, qui a été menée en 1982, a révélé qu'un inventaire ou un catalogue collectif des fichiers de données ordinologiques, sous forme automatisée ou imprimée, constituait de loin l'outil le plus en demande. À la réunion de consultation qui a eu lieu en juin 1983 à Vancouver, et au cours des nombreuses entrevues qui donneront lieu à la rédaction du rapport sur le « patri-moine électronique du Canada », la nécessité de dresser un inventaire a été soulignée maintes fois. Des chercheurs étaient d'avis que l'inventaire devait être établi par un organisme disposant d'une aide financière permanente et être à une plus petite échelle. Si de nombreux répertoires spécialisés ont été constitués depuis quelques années, il n'en demeure pas moins que leur portée est restreinte en comparaison de celle du Centre d'échange de données en sciences sociales.

L'aide à la recherche en sciences sociales a diminué dans les années 80, il devient nécessaire de connaître et d'utiliser les documents existants pour réduire le coût de la recherche. Une quantité énorme de données sont produites, données qui pourraient être utilisées si elles étaient identifiées. Les éléments descriptifs des données ordinales ont été normalisés et les programmes de gestion de l'information sont plus aptes à accepter des données sous d'autres formes. Ces facteurs facilitent la création d'un inventaire moins coûteux et plus utile. Les Archives ordinales étudient sérieusement la possibilité d'élaborer un projet pilote en vue de déterminer le coût et la rentabilité d'un tel inventaire. La dizaine d'archives qui y participeraient présenteraient à un centre des descriptions de leurs fonds. Ce centre convertirait les données sous un format standard ou, si les descriptions n'étaient pas ordinales, les enregistrerait. Un inventaire imprimé disponible. La première tâche serait d'en arriver à une entente sur les éléments descriptifs à utiliser. Il faudrait ensuite étudier le coût et la rentabilité

Campbell, alors bibliothécaire en chef des Toronto Public Libraries, entreprit une étude intitulée « Centre d'échange de données en sciences sociales » sur l'état des données ordinaires au début des années 70. L'étude comportait aussi un inventaire sommaire des fonds de divers centres de données, ainsi qu'un relevé de cent cinquante-sept universités et créateurs de données dans les universités canadiennes. Même s'ils ne créaient ni n'utilisaient pas tous des données ordinaires, la plupart des enseignants préoyaient en faire un usage assez considérable au cours des cinq années suivantes. M. Campbell conclut que la création et l'utilisation des données ordinaires étaient manifestes dans toutes les provinces et dans tous les secteurs. Il reconnaissait aussi l'importance des nombreuses bourses accordées par le Conseil des arts du Canada dans le domaine des sciences sociales, nombre d'entre elles ayant contribué à la création de données. Le comité directeur mixte AUCC/CRSHC accepta sa recommandation concernant l'établissement d'un centre d'échange de données en sciences sociales. Les principaux objectifs du centre étaient de 1) constituer un index de fonds ordinaires en sciences sociales dans les universités, les centres de recherche à but non lucratif et les autres organismes s'occupant de recherches en sciences sociales au Canada; 2) rassembler des notices de données ordinaires provenant des gouvernements fédéral et provinciaux et assurer la liaison entre les chercheurs et les ministères gouvernementaux; 3) renseigner les chercheurs sur l'emplacement des données; et 4) fournir des renseignements techniques pour assurer l'utilisation efficace des données.

C'est à la lumière de cette recommandation que fut créé le Centre d'échange de données en sciences sociales. L'une de ses tâches premières était de mettre au point une formule de description normalisée pour la collecte des renseignements sur les fichiers de données ordinologiques. Le Centre encouragea les archivistes canadiens à proposer les éléments à inclure dans la description. Au cours des années qui suivirent, le Centre rassembla des renseignements précieux sur l'emplacement et le contenu des données ordinologiques utilisées dans les universités et dans les secteurs public et privé. Une base de données importante fut créée et un inventaire fut publié et distribué. Le Centre publia aussi un bulletin qui décrivait les activités des archives, tant sur le plan national qu'international; les principaux projets de

Depuis la fin des années 60, deux préoccupations majeures occupent les archives et les bibliothèques du Canada. *Primo*, assurer l'intégralité de la documentation et la normalisation des codes; *secundo*, dresser un inventaire complet des données ordinolingues disponibles pour la recherche secondaire. Si des tentatives ont été faites afin de créer un inventaire, il a été impossible d'en assurer la tenue, par manque de ressources financières. L'établissement d'un inventaire est d'une grande importance, tant pour les archivistes que pour les chercheurs qui œuvrent dans le domaine des sciences sociales. Nombre de chercheurs doivent se servir de données américaines parce que les données canadiennes sont difficilement accessibles. L'absence d'information sur les sources de données ordinolingues au Canada entraîne un double emploi et des dépenses inutiles. Les ordinateurs furent utilisés pour la première fois à des fins statistiques au cours des

sciences sociales des universités pour créer et manipuler des données socio-économiques. Statistique Canada est depuis toujours le principal producteur de données informatiques. D'autres ministères fédéraux sont toutefois devenus d'importants producteurs de données, soit par l'entremise de services de sous-traitance. La Commission royale d'enquête sur le bilinguisme et le biculturalisme avait mené plusieurs enquêtes nationales par le biais de contrats octroyés à des chercheurs en sciences sociales. Les universités créèrent des archives pour acquérir, conserver et diffuser des séries de données aux universitaires. L'Institut for Behavioural Research ouvrit un centre de recherche avancée qui, en 1971, fut organisé en trois secteurs spécialisés : la Canadian Attitude Archive, la Cross-national Archive et la Census Data Archive. L'université de la Saskatchewan exploita le Sample Survey Centre et la Data Bank, et l'université de Montréal, le Centre de sondage. Même si elles étaient petites, les bibliothèques et les archives constituaient des centres vitaux de dissémination des données ordinolingués.

Le manque d'information sur les sources de données canadiennes a été reconnu dans un rapport rédigé pour le compte du Comité directeur mixte de l'Association des universités de recherches en sciences humaines du Canada (CRSHC) au début des années 70. M. H.C.